Расчетно-аналитическая работа
Анализ данных
Финансовый университет
Вариант 1

Задача 1.

В отделении Сбербанка микрорайона пользуются банкоматом 20% населения из близлежащих домов. Какова вероятность того, что из 500 наудачу выбранных жителей микрорайона в этом отделении Сбербанка пользуются банкоматом:

а) 90 человек;

б) от 80 до 130 человек;

в) более 120 человек?

Решение.

а) Используем локальную теорему Муавра-Лапласа.

Вероятность того, что в серии из n независимых опытов, в каждом из которых событие наступает с вероятностью p, событие наступит ровно k раз, приближенно равна

$$P_n(k) \approx \frac{1}{\sqrt{npq}}\varphi\left(\frac{k-np}{\sqrt{npq}}\right)$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

Значения $\varphi(x)$ находятся из таблиц (функция четная).

Имеем $n = 500, p = 0.2, k = 90$, получим:

$$P_{500}(90) \approx \frac{1}{\sqrt{500\cdot 0.2\cdot 0.8}}\varphi\left(\frac{90-500\cdot 0.2}{\sqrt{500\cdot 0.2\cdot 0.8}}\right) = \frac{1}{\sqrt{80}}\varphi\left(\frac{-10}{\sqrt{80}}\right) \approx 0.111803\cdot\varphi(1.12) \approx$$
$$\approx 0.111803\cdot 0.2131 \approx 0.0238$$

б) Используем интегральную теорему Муавра-Лапласа.

Вероятность того, что в серии из n независимых испытаний, в каждом из которых событие происходит с вероятностью p, событие произойдет от k_1 до k_2 раз, приближенно находится с помощью интегральной теоремы Муавра-Лапласа:

$$P_n(k_1;k_2) \approx \Phi\left(\frac{k_2-np}{\sqrt{npq}}\right) - \Phi\left(\frac{k_1-np}{\sqrt{npq}}\right)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-z^2/2}dz$$

$\Phi(x)$ - функция Лапласа, ее значения находятся из таблиц, с учетом того, что эта функция - нечетная.

Имеем $n = 500, p = 0.2, k_1 = 80, k_2 = 130$, получим:

$$P_{500}\,(80;130) \approx \Phi\left(\frac{130-500\cdot 0.2}{\sqrt{500\cdot 0.2\cdot 0.8}}\right) - \Phi\left(\frac{80-500\cdot 0.2}{\sqrt{500\cdot 0.2\cdot 0.8}}\right) = \Phi\left(\frac{30}{\sqrt{80}}\right) - \Phi\left(\frac{-20}{\sqrt{80}}\right) \approx$$
$$\approx \Phi(3.35) + \Phi(2.24) \approx 0.499596 + 0.4875 \approx 0.9871$$

в) Используем интегральную теорему Муавра-Лапласа.

$$P_{500}(k > 120) = 1 - P_{500}(0; 120)$$

Имеем $n = 500, p = 0.2, k_1 = 0, k_2 = 120$, получим:

$$P_{500}(0; 120) \approx \Phi\left(\frac{120 - 500 \cdot 0.2}{\sqrt{500 \cdot 0.2 \cdot 0.8}}\right) - \Phi\left(\frac{0 - 500 \cdot 0.2}{\sqrt{500 \cdot 0.2 \cdot 0.8}}\right) = \Phi\left(\frac{20}{\sqrt{80}}\right) - \Phi\left(\frac{-100}{\sqrt{80}}\right) \approx$$

$$\approx \Phi(2.24) + \Phi(11.18) \approx 0.4875 + 0.5 = 0.9875$$

$$P_{500}(k > 120) \approx 1 - 0.9875 = 0.0125$$

Ответ.

а) ≈ 0.0238; б) ≈ 0.9871; в) ≈ 0.0125

Задача 2.

По наблюдениям за температурой воздуха в сентябре этого года в данной местности установлено, что средняя температура воздуха составила 15°С, а среднее квадратическое отклонение равно 5°С. Оценить вероятность того, что в сентябре следующего года средняя температура воздуха будет:

а) не более 25℃;

б) более 20℃;

в) будет отличаться от средней температуры этого года не более чем на 7℃ (по абсолютной величине);

г) будет отличаться от средней температуры этого года не менее чем на 8℃ (по абсолютной величине).

Решение.

а) Используем неравенство Маркова:

$$P(X \le b) > 1 - \frac{M(X)}{b}$$

Учитывая, что $M(X) = 15, \sigma(X) = 5$, получим:

$$P(X \le 25) > 1 - \frac{15}{25}$$

$$P(X \le 25) > 0.4$$

б) Используем неравенство Маркова:

$$P(X > a) \le \frac{M(X)}{a}$$

$$P(X > 20) \le \frac{15}{20}$$

$$P(X > 20) \le 0.75$$

в) Используем неравенство Чебышева:

$$P(|X - M(X)| \le \varepsilon) > 1 - \frac{D(X)}{\varepsilon^2}$$

$$P(|X - M(X)| \le 7) > 1 - \frac{5^2}{7^2}$$

$$P(|X - M(X)| \le 7) > \frac{24}{49}$$

г) Используем неравенство Чебышева:

$$P(|X - M(X)| \geq \varepsilon) \leq \frac{D(X)}{\varepsilon^2}$$

$$P(|X - M(X)| \geq 8) \leq \frac{5^2}{8^2}$$

$$P(|X - M(X)| \geq 8) \leq \frac{25}{64}$$

Ответ.

а) $P(X \leq 25) > 0.4$; б) $P(X > 20) \leq 0.75$;

в) $P(|X - M(X)| \leq 7) > \frac{24}{49}$; г) $P(|X - M(X)| \geq 8) \leq \frac{25}{64}$

Задача 3.

Известно, что месячная доходность некоторой ценной бумаги есть нормально распределенная случайная величина ξ. Найти ее математическое ожидание и среднее квадратическое отклонение, если известно, что $P(\xi < 1) = 0.1$ и $P(\xi \geq 5) = 0.5$.

Построить схематично графики функции распределения и функции плотности распределения этой случайной величины.

Вычислить вероятность того, что в следующем месяце доходность ценной бумаги будет:

а) не более 4%;

б) не менее 8%;

в) от 3% до 7%.

Решение.

Для нормально распределенной случайной величины с параметрами a, σ вероятность попадания в интервал $(\alpha; \beta)$ находится по формуле:

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right)$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2} dz$$

$\Phi(x)$ - функция Лапласа, ее значения находятся из таблиц, эта функция - нечетная.

Учитывая, что нормально распределенная случайная величина симметрична относительно математического ожидания, $P\big(X \geq M(X)\big) = P\big(X \leq M(X)\big) = 0.5$. Таким образом, так как $P(\xi \geq 5) = 0.5$, то $M\xi = 5$. С учетом этого,

$$P(\xi < 1) = \Phi\left(\frac{1 - 5}{\sigma}\right) - \Phi\left(\frac{-\infty - 5}{\sigma}\right) = \Phi\left(\frac{-4}{\sigma}\right) - \Phi(-\infty) = -\Phi\left(\frac{4}{\sigma}\right) - (-0.5) =$$

$$= 0.5 - \Phi\left(\frac{4}{\sigma}\right) = 0.1$$

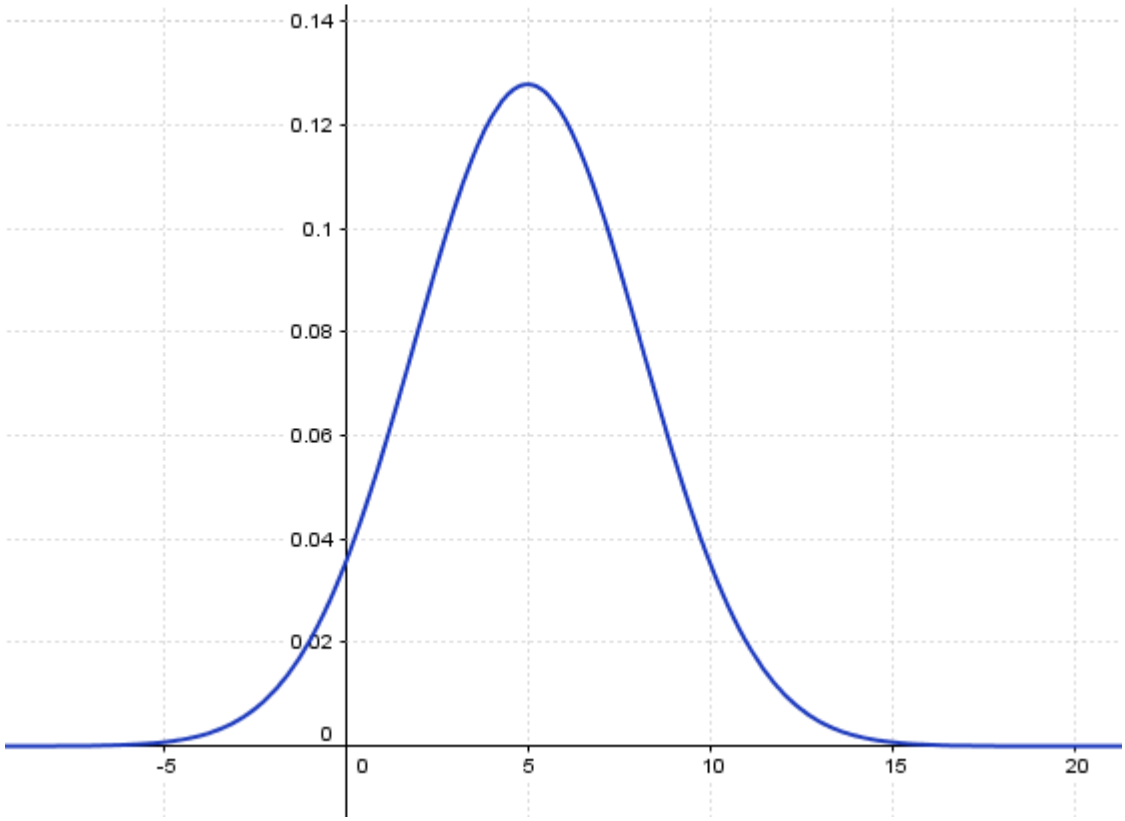
$$\Phi\left(\frac{4}{\sigma}\right) = 0.5 - 0.1 = 0.4$$

По таблице значений функции Лапласа найдем

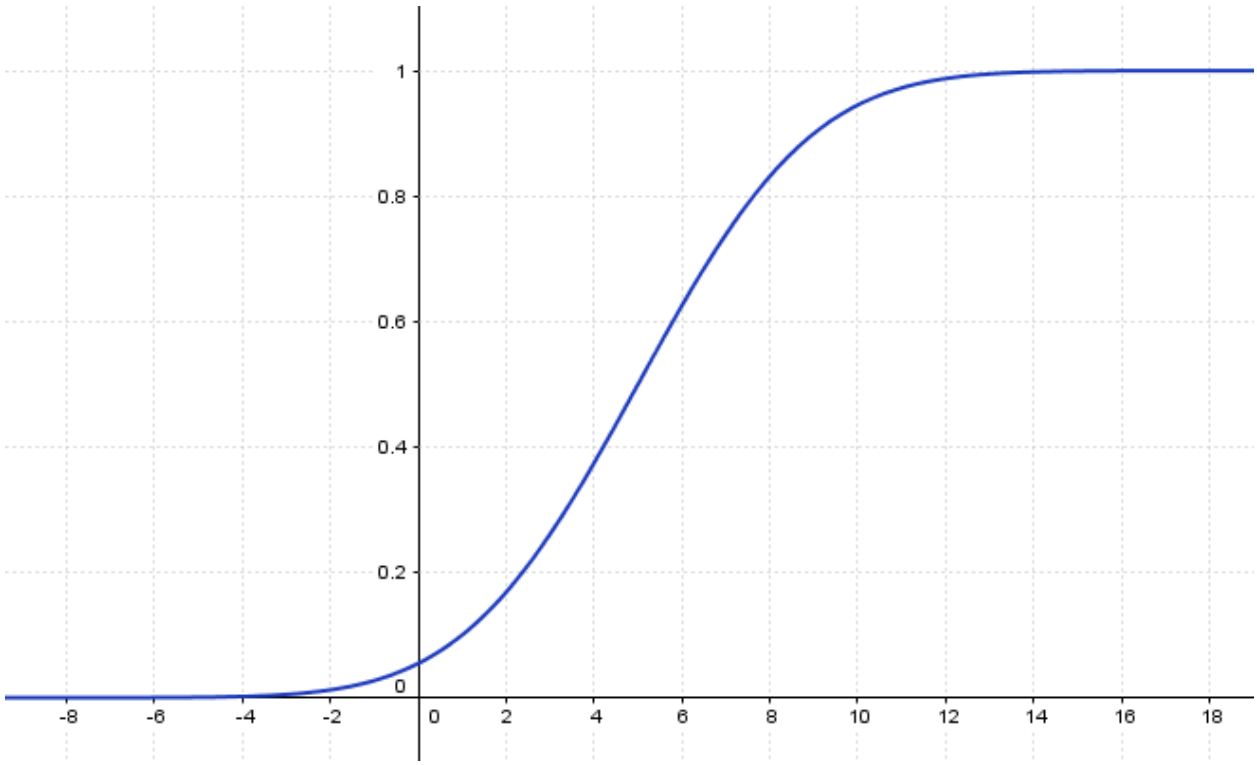
$$\Phi(1.2816) \approx 0.4 \rightarrow \frac{4}{\sigma} \approx 1.2816 \rightarrow \sigma \approx \frac{4}{1.2816} \approx 3.1212$$

Построим схематично графики функции распределения и функции плотности распределения этой случайной величины.

Плотность распределения:



Функция распределения:



а) Вероятность того, что в следующем месяце доходность ценной бумаги будет не более 4%:

$$P(\xi \le 4) = \Phi\left(\frac{4-5}{3.1212}\right) - \Phi\left(\frac{-\infty-5}{3.1212}\right) \approx \Phi(-0.32) - \Phi(-\infty) \approx -0.1255 + 0.5 = 0.3745$$

б) Вероятность того, что в следующем месяце доходность ценной бумаги будет не менее 8%:

$$P(\xi \ge 8) = \Phi\left(\frac{+\infty-5}{3.1212}\right) - \Phi\left(\frac{8-5}{3.1212}\right) \approx \Phi(+\infty) - \Phi(0.96) \approx 0.5 - 0.3315 = 0.1685$$

в) Вероятность того, что в следующем месяце доходность ценной бумаги будет от 3% до 7%:

$$P(3 \le \xi \le 4) = \Phi\left(\frac{4-5}{3.1212}\right) - \Phi\left(\frac{3-5}{3.1212}\right) \approx \Phi(-0.32) - \Phi(-0.64) \approx -0.1255 + 0.2389 = 0.1134$$

Ответ.

$M(\xi) = 5, \sigma(\xi) \approx 3.1212;$

а) ≈ 0.3745; б) ≈ 0.1685; в) ≈ 0.1134

Задача 4.

С целью изучения миграции населения в данной области было проведено выборочное обследование 70 мелких населенных пунктов из 350 имеющихся в области (выборка бесповторная). Получены следующие данные о количестве зарегистрированных мигрантов:

9	0	8	3	10	5	14	6	14	1
3	4	10	5	4	11	4	14	13	13
12	2	1	3	9	14	0	10	5	7
3	11	6	3	14	7	2	2	6	10
8	5	9	14	7	7	0	3	11	7
12	13	2	13	5	14	6	13	3	1
6	8	9	7	5	13	13	7	1	12

Составить интервальный вариационный ряд. Записать эмпирическую функцию распределения и построить ее график. На одном чертеже изобразить гистограмму и полигон частот.

По сгруппированным данным вычислить выборочные числовые характеристики: среднее арифметическое, исправленную выборочную дисперсию, среднее квадратичное отклонение, коэффициент вариации, асимметрию, эксцесс, моду и медиану. Найти:

а) вероятность того, что среднее количество мигрантов во всей области отличается от их среднего количества в выборке не более чем на 1 чел;

б) границы, в которых с вероятностью 0,98 заключена доля всех населенных пунктов области, где количество мигрантов превышает 8 человек;

в) объем бесповторной выборки, при котором те же границы для среднего количества мигрантов, что и в п. а) можно гарантировать с вероятностью 0,95.

Решение.

Объем выборки $n = 70, x_{min} = 0, x_{max} = 14$. Определим оптимальное число интервалов по формуле Стерджесса:

$$1 + \log_2 n = 1 + \log_2 70 \approx 7.13$$

Округляя до ближайшего целого, получим $k = 7$ интервалов. Длина интервала:

$$l = \frac{x_{max} - x_{min}}{k} = \frac{14 - 0}{7} = 2$$

Разобьем отрезок $[0; 14]$ на интервалы $[0; 2]$, $(2; 4]$, $(4; 6]$, … , $(12; 14]$. Подсчитаем частоты попадания значений выборки в интервалы. Построим интервальный вариационный ряд:

Границы интервала		Частота
0	2	11
2	4	10
4	6	11
6	8	10
8	10	8
10	12	6
12	14	14
	Сумма	70

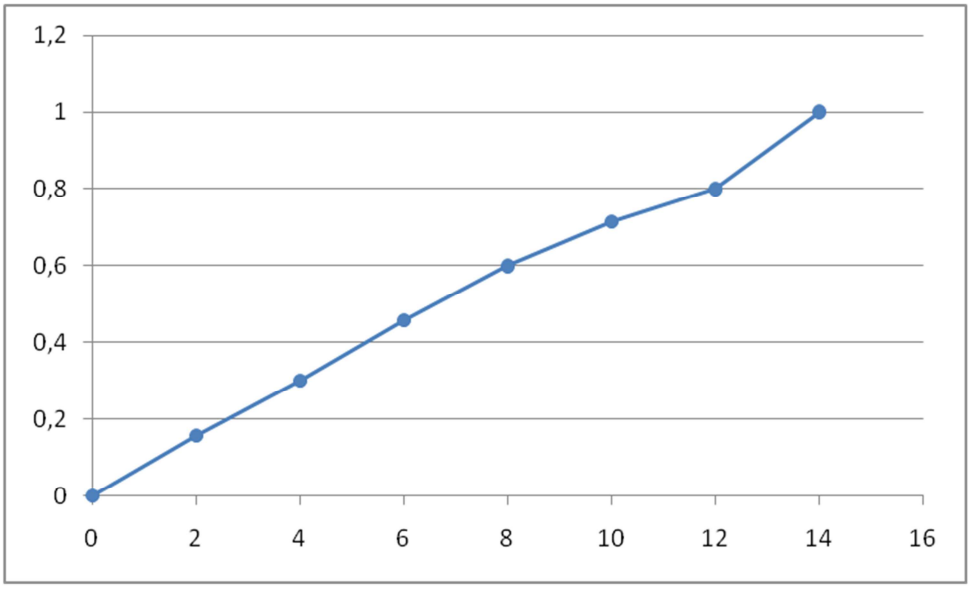
Найдем относительные частоты $w_i = \frac{n_i}{n}$, накопленные относительные частоты $w_i^{\text{нак}}$:

Интервал		n_i	w_i	$w_i^{\text{нак}}$
0	2	11	0,1571	0,1571
2	4	10	0,1429	0,3000
4	6	11	0,1571	0,4571
6	8	10	0,1429	0,6000
8	10	8	0,1143	0,7143
10	12	6	0,0857	0,8000
12	14	14	0,2000	1,0000

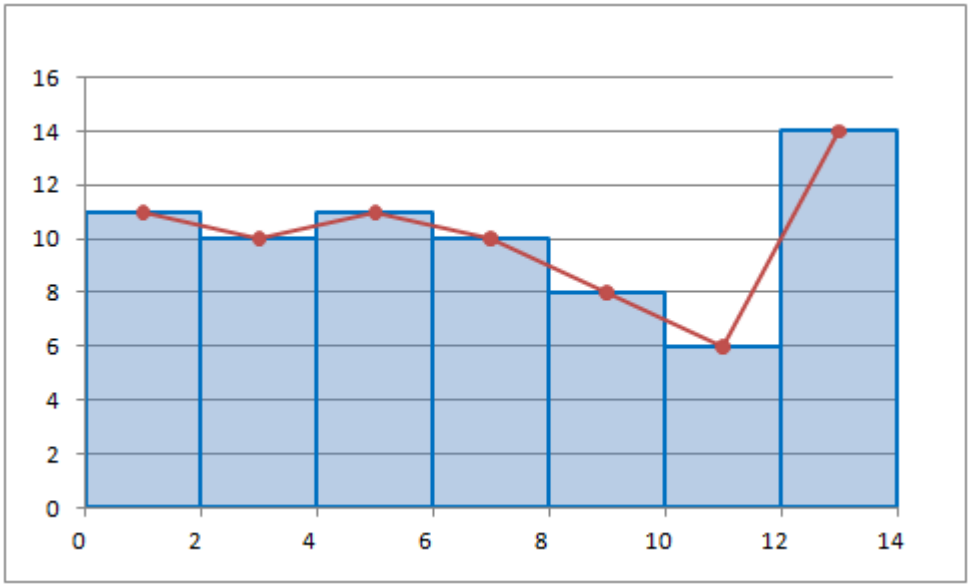
Запишем эмпирическую функцию распределения:

x_i	$F^*(x_i)$
0	0
2	0,1571
4	0,3000
6	0,4571
8	0,6000
10	0,7143
12	0,8000
14	1,0000

Так как таблица определяет функцию не полностью, то при изображении графика доопределяем функцию, соединяя точки графика, соответствующие концам интервалов, отрезками.



Построим гистограмму и полигон частот на одном графике:



По сгруппированным данным вычислим выборочные числовые характеристики: среднее арифметическое, исправленную выборочную дисперсию, среднее квадратичное отклонение, коэффициент вариации, асимметрию, эксцесс, моду и медиану.
Промежуточные расчеты приведем в таблице:

Интервал		x_i	n_i	$x_i n_i$	$x_i^2 n_i$
0	2	1	11	11	11
2	4	3	10	30	90
4	6	5	11	55	275
6	8	7	10	70	490
8	10	9	8	72	648
10	12	11	6	66	726

12	14	13	14	182	2366
			Сумма	486	4606

Среднее арифметическое:

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{486}{70} = 6.9429$$

Исправленная выборочная дисперсия:

$$S^2 = \frac{n}{n-1}\left(\frac{\sum x_i^2 n_i}{n} - \bar{x}^2\right) = \frac{70}{69}\left(\frac{4606}{70} - 6.9429^2\right) = 17.8518$$

Среднее квадратичное отклонение:

$$s = \sqrt{S^2} = \sqrt{17.8518} = 4.2251$$

Коэффициент вариации:

$$V = \frac{s}{\bar{x}} \cdot 100\% = \frac{4.2251}{6.9429} \cdot 100\% = 60.86\%$$

Промежуточные вычисления для нахождения асимметрии и эксцесса:

x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3$	$(x_i - \bar{x})^3 n_i$	$(x_i - \bar{x})^4$	$(x_i - \bar{x})^4 n_i$
1	11	-5,9429	-209,8872	-2308,7588	1247,3294	13720,6235
3	10	-3,9429	-61,2961	-612,9614	241,6819	2416,8192
5	11	-1,9429	-7,3337	-80,6706	14,2483	156,7315
7	10	0,0571	0,0002	0,0019	0,0000	0,0001
9	8	2,0571	8,7055	69,6439	17,9084	143,2675
11	6	4,0571	66,7822	400,6934	270,9450	1625,6702
13	14	6,0571	222,2304	3111,2255	1346,0812	18845,1371
			Сумма	579,1739		36908,2492

Центральный момент третьего порядка и асимметрия:

$$\mu_3 = \frac{\sum (x_i - \bar{x})^3 n_i}{n} = \frac{579.1739}{70} = 8.2739; \;\; A_s = \frac{\mu_3}{S^3} = \frac{8.2739}{4.2251^3} = 0.1097$$

Центральный момент четвертого порядка и эксцесс:

$$\mu_4 = \frac{\sum (x_i - \bar{x})^4 n_i}{n} = \frac{36908.2492}{70} = 527.2607; \;\; Ex = \frac{\mu_4}{S^4} - 3 = \frac{527.2607}{4.2251^4} - 3 = 112.8141$$

Мода представляет собой значение изучаемого признака, повторяющегося с наибольшей частотой. Расчет моды для интервальных вариационных рядов производится по формуле:

$$Mo = x_0 + h \cdot \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})}$$

где x_0 - начало модального интервала (то есть интервала с наибольшей частотой), h - длина интервала, n_{Mo} - частота модального интервала, n_{Mo-1} - частота интервала, предшествующего модальному, n_{Mo+1} – частота интервала, следующего за модальным. Получим:

$$Mo = 12 + 2 \cdot \frac{14 - 6}{(14 - 6) + (14 - 0)} = 12.7273$$

Медианой называется значение признака, приходящегося на середину ранжированной (упорядоченной) совокупности. Расчет медианы для интервальных вариационных рядов производится по формуле:

$$Me = x_0 + h \cdot \frac{\frac{1}{2}n - S_{Me-1}}{n_{Me}}$$

где x_0 - нижняя граница медианного интервала (медианным называется первый интервал, накопленная частота которого превышает половину общей суммы частот); n - сумма всех частот ряда; h - величина медианного интервала; S_{Me-1} - накопленная частота интервала, предшествующего медианному; n_{Me}- частота медианного интервала. Первый интервал, накопленная относительная частота которого больше 0.5, - $(6; 8]$. Получим:

$$Me = 6 + 2 \cdot \frac{35 - (11 + 10 + 11)}{10} = 6.6$$

а) Найдем вероятность того, что среднее количество мигрантов во всей области отличается от их среднего количества в выборке не более чем на 1 чел.

$$P(|\bar{X}_\text{г} - \bar{x}| \le \Delta) = 2\Phi(t) = \gamma$$

$$t = \frac{\Delta}{\sigma_x}$$

Стандартная ошибка бесповторной выборки для средней:

$$\sigma_x = \sqrt{\frac{S^2}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{17.8518}{70}\left(1 - \frac{70}{350}\right)} = 0.4571$$

$$t = \frac{\Delta}{\sigma_x} = \frac{1}{0.4571} = 2.2139$$

$$\gamma = 2\Phi(t) = 2\Phi(2.2139) \approx 2 \cdot 0.4866 = 0.9732$$

б) Найдем границы, в которых с вероятностью 0,98 заключена доля всех населенных пунктов области, где количество мигрантов превышает 8 человек.

Выборочная доля населенных пунктов области, где количество мигрантов превышает 8 человек:

$$w = \frac{m}{n} = \frac{8 + 6 + 14}{70} = 0.4$$

Стандартная ошибка выборочной доли:

$$\sigma_w = \sqrt{\frac{w(1-w)}{n}\left(1 - \frac{n}{N}\right)} = \sqrt{\frac{0.4 \cdot (1 - 0.4)}{70}\left(1 - \frac{70}{350}\right)} = 0.0580$$

Из условия $\Phi(t) = \frac{\gamma}{2}$ найдем t

$$\Phi(t) = \frac{\gamma}{2} = \frac{0.98}{2} = 0.49; \;\; t = 2.3263$$

$$\Delta = \sigma_w t = 0.0580 \cdot 2.3263 = 0.1348$$

Итак, искомый интервал: $(0.4 - 0.1348; 0.4 + 0.1348) = (0.2652; 0.5348)$

в) Найдем объем бесповторной выборки, при котором те же границы для среднего количества мигрантов, что и в п. а) можно гарантировать с вероятностью 0,95.

Определим t

$$\Phi(t) = \frac{\gamma}{2} = \frac{0.95}{2} = 0.475;\ \ t = 1.9600$$

Так как границы должны остаться теми же, то $\Delta = 1$, тогда

$$n' = \frac{Nt^2s^2}{t^2s^2 + N\Delta^2} = \frac{350 \cdot 1.96^2 \cdot 17.8518}{1.96^2 \cdot 17.8518 + 350 \cdot 1^2} = 57.3416$$

Итак, потребуется объем выборки $n \geq 58$.

Ответ.

$\bar{x} = 6.9429;\ \ S^2 = 17.8518;\ \ s = 4.2251; V = 60.86\%$

$As = 0.1097;\ \ Ek = 112.8141;\ \ Mo = 12.7273;\ \ Me = 6.6$

а) ≈ 0.9732; б) $(0.2652; 0.5348)$; в) $n \geq 58$.

Задача 5.

Заменив неизвестные параметры генеральной совокупности соответственно их наилучшими выборочными оценками, по данным задачи 4, используя χ^2-критерий Пирсона на уровне значимости $\alpha = 0.05$, проверить две гипотезы о том, что изучаемая случайная величина ξ - число мигрантов в данном населенном пункте - распределена:

а) по нормальному закону распределения;

б) по равномерному закону распределения.

Построить на чертеже, где изображена гистограмма эмпирического распределения, соответствующие графики равномерного и нормального распределений.

Решение.

Заменим неизвестные параметры генеральной совокупности их наилучшими выборочными оценками, то есть положим

$$M(\xi) = \bar{x} = 6.9429;\ \ D(\xi) = S^2 = 17.8518$$

а) Проверим по критерию Пирсона гипотезу о том, что изучаемая случайная величина ξ - число мигрантов в данном населенном пункте - распределена по нормальному закону с параметрами $a = 6.9429, \sigma = 4.2251$

Найдем теоретические частоты попадания в интервалы. Нормируем границы интервалов:

$$z_i = \frac{x_i - a}{\sigma} = \frac{x_i - 6.9429}{4.2251}$$

Началом первого интервала положим $-\infty$, концом последнего $+\infty$. Определим теоретические вероятности попадания случайной величины в интервалы: $P_i = \Phi(z_{i+1}) - \Phi(z_i)$ и теоретические частоты: $n_i' = nP_i = 70P_i$. Получим:

x_i	x_{i+1}	z_i	$z_{(i+1)}$	$\Phi(z_i)$	$\Phi(z_{i+1})$	P_i	n_i'
0	2	$-\infty$	-1,1699	-0,5000	-0,3790	0,1210	8,472
2	4	-1,1699	-0,6965	-0,3790	-0,2569	0,1220	8,542

4	6	-0,6965	-0,2232	-0,2569	-0,0883	0,1687	11,806
6	8	-0,2232	0,2502	-0,0883	0,0988	0,1871	13,095
8	10	0,2502	0,7236	0,0988	0,2653	0,1665	11,658
10	12	0,7236	1,1969	0,2653	0,3843	0,1190	8,330
12	14	1,1969	$+\infty$	0,3843	0,5000	0,1157	8,097

Вычислим наблюдаемое значение критерия Пирсона:

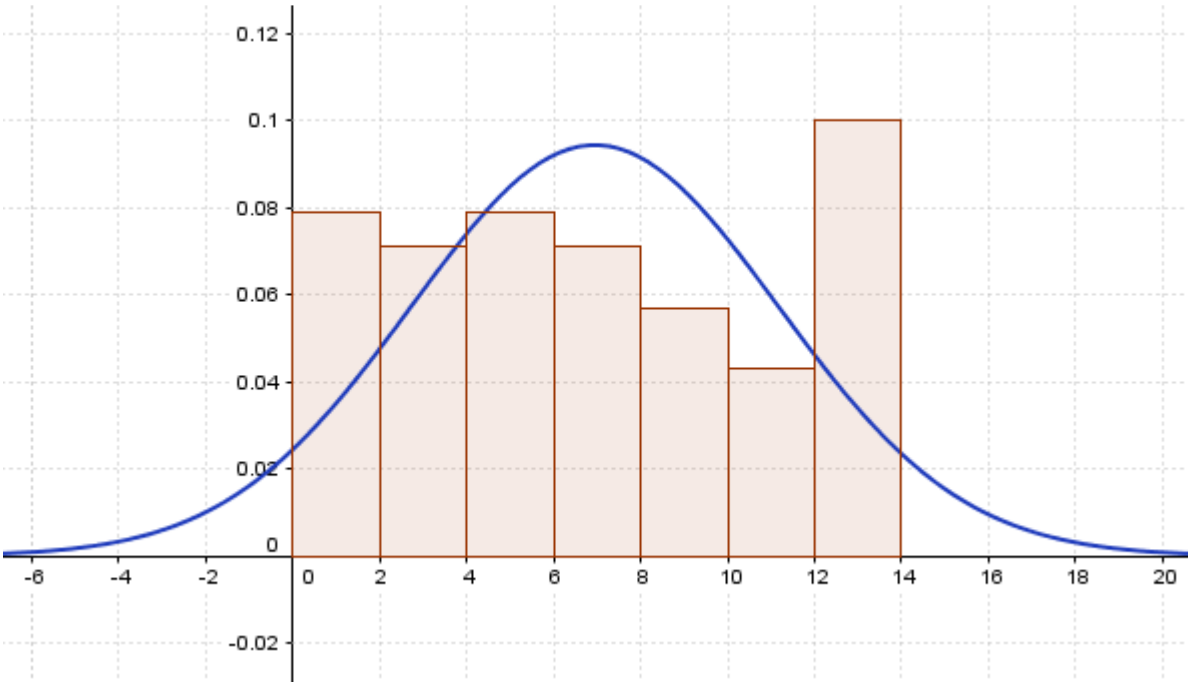
$$\chi^2_{\text{набл}} = \sum \frac{(n_i - n_i')^2}{n_i'}$$

n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
11	8,472	2,528	6,3915	0,7544
10	8,542	1,458	2,1260	0,2489
11	11,806	-0,806	0,6492	0,0550
10	13,095	-3,095	9,5816	0,7317
8	11,658	-3,658	13,3833	1,1480
6	8,330	-2,330	5,4284	0,6517
14	8,097	5,903	34,8473	4,3038
			Сумма	7,8935

$$\chi^2_{\text{набл}} = 7.8935$$

По таблице критических точек распределения χ^2 по уровню значимости $\alpha = 0.05$ и числу степеней свободы $k = s - 3 = 7 - 3 = 4$ найдем $\chi^2_{\text{крит}} = 9.4877$. Так как $\chi^2_{\text{набл}} < \chi^2_{\text{крит}}$, нет основания отвергнуть гипотезу о нормальном распределении случайной величины ξ.

Построим на одном графике гистограмму относительных частот (прямоугольники с основаниями $[x_i; x_{i+1}]$ и высотой $h_i = \frac{w_i}{l} = \frac{w_i}{2}$) и плотность теоретического нормального распределения с параметрами $a = 6.9429, \sigma = 4.2251$:



б) Проверим по критерию Пирсона гипотезу о том, что изучаемая случайная величина ξ - число мигрантов в данном населенном пункте - распределена по равномерному закону.

Оценки параметров a, b нормального распределения найдем методом моментов, зная, что $M(\xi) = 6.9429;\ \ D(\xi) = 17.8518$. Получим:

$$\begin{cases} M(\xi) = \dfrac{a+b}{2} \\ D(\xi) = \dfrac{(b-a)^2}{12} \end{cases} \rightarrow \begin{cases} b + a = 2M(\xi) \\ b - a = 2\sqrt{3}\sigma(\xi) \end{cases} \rightarrow \ b = M(\xi) + \sqrt{3}\sigma(\xi);\ \ a = M(\xi) - \sqrt{3}\sigma(\xi)$$

$$a = 6.9429 - \sqrt{3 \cdot 17.8518} = -0.3753$$

$$b = 6.9429 + \sqrt{3 \cdot 17.8518} = 14.2610$$

Найдем теоретические частоты попадания в интервалы:

$$P_1 = P(\xi < x_1) = \frac{x_1 - a}{b - a};$$

$$P_i = P(x_i < \xi < x_{i+1}) = \frac{x_{i+1} - x_i}{b - a};\ \ i = 2, \dots, 6$$

$$P_7 = \frac{b - x_7}{b - a}$$

$$n_i' = nP_i = 70P_i$$

Получим:

x_i	x_{i+1}	P_i	n_i'
0	2	0,1623	11,3601
2	4	0,1366	9,5653
4	6	0,1366	9,5653
6	8	0,1366	9,5653
8	10	0,1366	9,5653
10	12	0,1366	9,5653
12	14	0,1545	10,8136

Вычислим наблюдаемое значение критерия Пирсона:

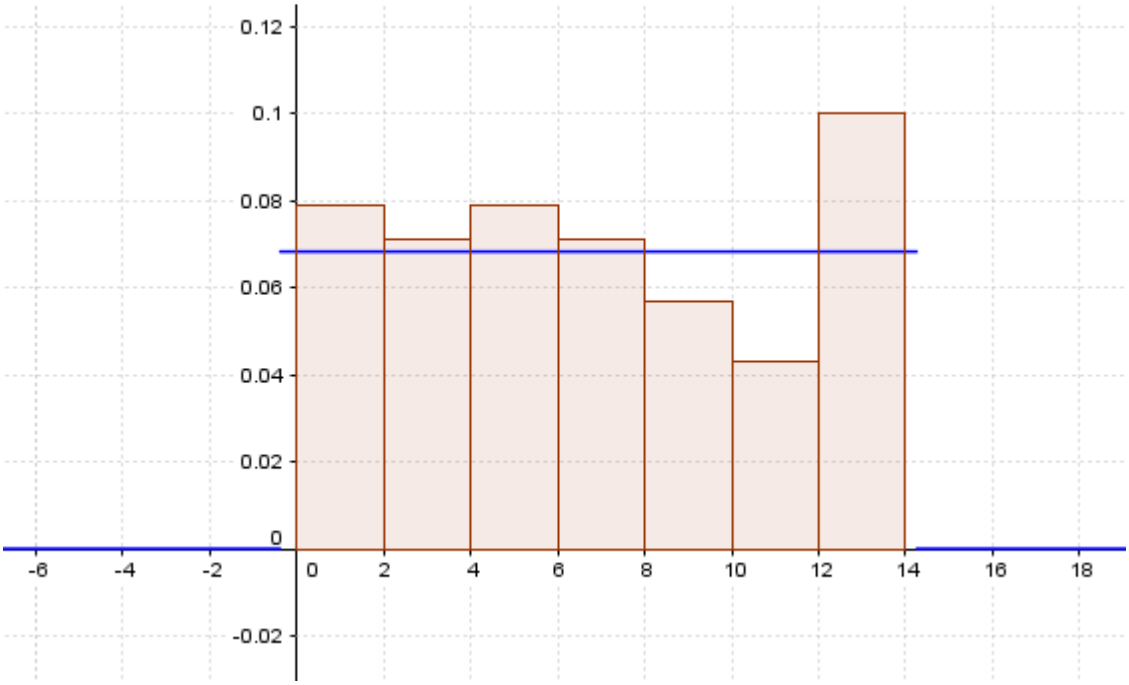
$$\chi^2_{\text{набл}} = \sum \frac{(n_i - n_i')^2}{n_i'}$$

n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
11	11,360	-0,360	0,1297	0,0114
10	9,565	0,435	0,1890	0,0198
11	9,565	1,435	2,0585	0,2152
10	9,565	0,435	0,1890	0,0198
8	9,565	-1,565	2,4500	0,2561
6	9,565	-3,565	12,7111	1,3289
14	10,814	3,186	10,1535	0,9390
			Сумма	2,7901

$$\chi^2_{\text{набл}} = 2.7901$$

По таблице критических точек распределения χ^2 по уровню значимости $\alpha = 0.05$ и числу степеней свободы $k = s - 3 = 7 - 3 = 4$ найдем $\chi^2_{\text{крит}} = 9.4877$. Так как $\chi^2_{\text{набл}} < \chi^2_{\text{крит}}$, нет основания отвергнуть гипотезу о равномерном распределении случайной величины ξ.

Теоретическая функция равномерного распределения на отрезке $[a; b]$ равна $\frac{1}{b-a} = \frac{1}{14.2610+0.3753} = 0.0683$ и равна 0 вне отрезка $[a; b]$. Построим на одном графике гистограмму относительных частот (прямоугольники с основаниями $[x_i; x_{i+1}]$ и высотой $h_i = \frac{w_i}{l} = \frac{w_i}{2}$) и плотность теоретического нормального распределения с параметрами $a = -0.3753, \sigma = 14.2610$:



Ответ.

На заданном уровне значимости нет оснований отвергнуть обе гипотезы о нормальном и равномерном распределении, однако равномерное распределение лучше согласуется с выборочными данными.

Задача 6.

С целью изучения зависимости количества времени использования клиентом мобильной связи в течение месяца ξ (мин) и стоимости минуты разговора η (руб.) произведено обследование 100 абонентов, пользующихся различными тарифными планами, и получены следующие данные:

ξ \ η	Менее 1	1-1,5	1,5-2	2-2,5	2,5-3	Более 3	Итого:
Менее 200				3	9	3	15
200-400				5	8	7	20
400-600			4	13	9	3	29
600-800		2	6	8	2		18
Более 800	6	5	6	1			18
Итого:	6	7	16	30	28	13	100

Необходимо:

1. Вычислить групповые средние $\bar{x}_i$ и $\bar{y}_j$, построить эмпирические линии регрессии.

2. Предполагая, что между переменными ξ и η существует линейная корреляционная зависимость:

а) найти уравнения прямых регрессии, построить их графики на одном чертеже с эмпирическими линиями регрессии и дать экономическую интерпретацию полученных уравнений;

б) вычислить коэффициент корреляции; на уровне значимости $\alpha = 0.05$ оценить его значимость и сделать вывод о тесноте и направлении связи между переменными ξ и η;
в) используя соответствующее уравнение регрессии, оценить время использования мобильной связи при стоимости минуты разговора 2,25 руб.

Решение.

Закроем открытые интервалы. Для ξ имеем: длина второго интервала равна 200, чтобы длина первого интервала была равна 200, положим началом первого интервала $200 - 200 = 0$. Длина предпоследнего интервала равна 200, концом последнего интервала положим $800 + 200 = 1000$. Аналогично для η началом первого интервала положим $1 - 0.5 = 0.5$, концом последнего $3 + 0.5 = 3.5$.
Для расчетов будем использовать середины интервалов:

	η						
ξ	0,75	1,25	1,75	2,25	2,75	3,25	Сумма
100				3	9	3	15
300				5	8	7	20
500			4	13	9	3	29
700		2	6	8	2		18
900	6	5	6	1			18
Сумма	6	7	16	30	28	13	

1. Вычислим групповые средние $\bar{x}_i$:

$$\bar{x}_1 = \frac{900 \cdot 6}{6} = 900$$

$$\bar{x}_2 = \frac{700 \cdot 2 + 900 \cdot 5}{7} = 842.8571$$

$$\bar{x}_3 = \frac{500 \cdot 4 + 700 \cdot 6 + 900 \cdot 6}{16} = 725$$

и так далее. Получим:

y_i	0,75	1,25	1,75	2,25	2,75	3,25
$\bar{x}_i$	900	842,8571	725	493,3333	328,5714	300

Аналогично,

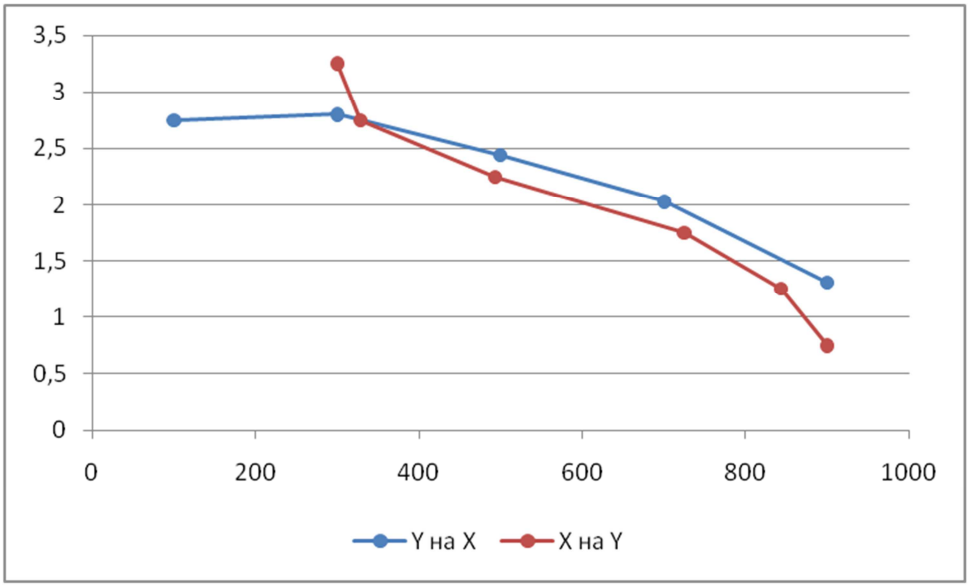
$$\bar{y}_1 = \frac{3 \cdot 2.25 + 9 \cdot 2.75 + 3 \cdot 3.25}{15} = 2.75$$

и так далее. Получим:

x_j	$\bar{y}_j$
100	2,75
300	2,8
500	2,4397
700	2,0278

900	1,3056

Построим эмпирические линии регрессии:



2. а) Найдем уравнения прямых регрессии.

Вычислим все необходимые величины.

x_i	n_{x_i}	$x_i n_{x_i}$	$x_i^2 n_{x_i}$
100	15	1500	150000
300	20	6000	1800000
500	29	14500	7250000
700	18	12600	8820000
900	18	16200	14580000
	Сумма	50800	32600000

$$\bar{x} = \frac{\sum x_i n_{x_i}}{n} = \frac{50800}{100} = 508$$

$$D_x = \overline{x^2} - \bar{x}^2 = \frac{32600000}{100} - 508^2 = 67936$$

$$\sigma_x = \sqrt{D_x} = \sqrt{67936} = 260.6454$$

y_i	n_{y_i}	$y_i n_{y_i}$	$y_i^2 n_{y_i}$
0,75	6	4,5	3,375
1,25	7	8,75	10,9375
1,75	16	28	49
2,25	30	67,5	151,875
2,75	28	77	211,75
3,25	13	42,25	137,3125
Сумма	100	228	564,25

$$\bar{y} = \frac{228}{100} = 2.28;\ \ D_y = \frac{564.25}{100} - 2.28^2 = 0.4441;\ \sigma_y = \sqrt{0.4441} = 0.664$$

Найдем $\overline{xy}$:

$$\overline{xy} = \frac{1}{100}\sum x_i y_j n_{ij} = \frac{103000}{100} = 1030$$

Запишем уравнение регрессии Y на X:

$$y = \bar{y} + \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{D_x}(x - \bar{x}) = 2.28 + \frac{1030 - 508\cdot 2.28}{67936}(x - 508)$$

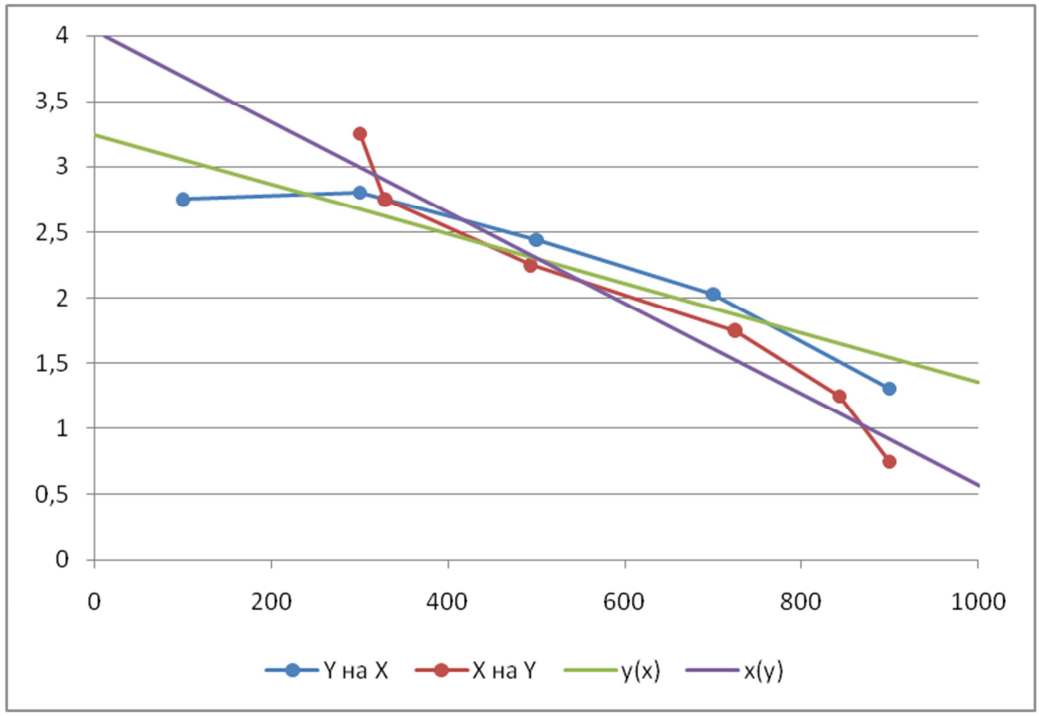
$$y = -0.0019x + 3.2389$$

Запишем уравнение регрессии X на Y:

$$x = \bar{x} + \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{D_y}(x - \bar{y}) = 508 + \frac{1030 - 508\cdot 2.28}{0.4441}(y - 2.28)$$

$$x = -288.7638y + 1166.3814$$

Построим графики прямых регрессии на одном чертеже с эмпирическими линиями регрессии



Уравнение $y = -0.0019x + 3.2389$ описывает (в среднем) зависимость стоимости минуты разговора (в рублях) от численности количества времени использования клиентом мобильной связи (в минутах). Выборочные данные позволяют предполагать, что с увеличением количества времени использования клиентом мобильной связи на минуту стоимость минуты разговора в среднем уменьшается на 0.0019 руб.

Уравнение $x = -288.7638y + 1166.3814$ описывает (в среднем) зависимость количества времени использования клиентом мобильной связи (в минутах) от стоимости минуты разговора (в рублях). Выборочные данные позволяют предполагать, что с увеличением стоимости минуты разговора (в рублях) на 1 рубль количество времени использования клиентом мобильной связи уменьшится в среднем на 288.76 минут.

б) Вычислим коэффициент корреляции.

$$r_{\text{в}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} = \frac{1030 - 508 \cdot 2.28}{260.6454 \cdot 0.6664} = -0.7383$$

На уровне значимости $\alpha = 0.05$ оценим его значимость. Выдвинем нулевую гипотезу о равенстве нулю генерального коэффициента корреляции и проверим ее.

$$T_{\text{набл}} = \frac{r_{\text{в}}\sqrt{n-2}}{\sqrt{1-r_{\text{в}}^2}} = \frac{-0.7383\sqrt{98}}{\sqrt{1-(-0.7383)^2}} = -10.8363$$

По таблице критических точек распределения Стьюдента по уровню значимости 0.05 и числу степеней свободы $k = n - 2 = 98$ найдем критическую точку двусторонней критической области

$$t_{\text{кр}} = 1.9845$$

Так как $|T_{\text{набл}}| > t_{\text{кр}}$, гипотезу о равенстве нулю генерального коэффициента корреляции следует отвергнуть, выборочный коэффициент корреляции значим.
Между переменными существует заметная обратная связь.

в) Используя уравнение регрессии $x = -288.7638y + 1166.3814$, оценим время использования мобильной связи при стоимости минуты разговора 2,25 руб.

$$x(y = 2.25) = -288.7638 \cdot 2.25 + 1166.3814 = 516.6629$$